GCC Tour Guide Manual

CD DocDB Document 2361-v7.1.
Last updated: Friday, June 14, 2013 by Ken Schumacher
February 2010 by Anne Heavey, Dave Ritchie;

This is currently a work in progress, based on document 2361-v7. Feedback and corrections are appreciated.

Introduction

Procedure for requesting a tour

When some group wishes to request a tour of Computing Sector facilities, we ask that they submit a request via the Service Desk. Specify that the request be assigned to the 'Outreach Support Group'. The leader or contact person for the tour group is asked to provide the following information:

- Describe the group and any affiliation they may have with Fermilab.
- What is the scope of the tour being requested?
  - Which facilities are you asking to visit: Grid Computing Center(GCC), Feynman Computing Center (FCC), or some other facility.
  - Is there any topic of special interest for the group, such as High Performance Computing, Grid computing, CMS computing facilities or Disk/tape storage.
- How many people, relative age of the audience and the technical level (science/computing) of the audience. For student tours, how many chaperones will there be. Groups of over 15 persons may need to be split into smaller groups.
- When do they want this tour (date/time).

Tour guides, please:

Ahead of time:

- Request access to GCC areas. See http://cdorg.fnal.gov/fop/CD_Controlled_Access.htm (Computing Sector Controlled Access)
- Complete your training (CD Computer Room Hazard Analysis) and access permission, as described on above Web page
- Review: http://cdorg.fnal.gov/fop/tour_instructions.html (Computer Room Tour Guide Safety Instructions)
- Read through this manual to become familiar with the tour information. Depending on the knowledge level and interests of the tour participants, you may want to emphasize (or de-emphasize) various points.
- Request/gather any handouts (literature, postcards) as needed. Contact CS Communications group (through service desk ticket) to get postcards and/or book marks. For brochures, contact the Office of Communication, x3351. Coordinate handouts with the leader who requested the tour.

When you arrive at GCC

- Sign in the visitor log book to register the tour group. There is a log book in each entry foyer.

Emphasize safety

Brief visitors either outside the building or in the entry foyer regarding safety. It is easier for them to hear the safety instructions while you are outside of our noisy rooms. Tell your tours:

- No food or drink allowed.
- No flammable items/substances.
- Don't touch any equipment or cords.
- Stay with tour guide and group.
- Possible hazards: Uneven floor tiles, ramps and steps, ongoing work areas, accidental equipment resets and unplugged cables. As you move between rooms, ask visitors to point out steps to those following behind them.
- Do not enter areas where yellow cones and/or yellow-chained barriers are in place.
- Avoid active work areas.
- Be especially cautious near areas where floor tiles have been removed. Do not stand with your back to open areas of the floor.
- Each of the equipment rooms is rather noisy. There is hearing protection available outside each computer room.
- One special note: when moving from CRA to the Battery room, there is a step. remind the visitors to point this out to the person behind them, so no one falls.

Keep areas clean

- A significant source of dirt and dust in the Data Centers comes from people's shoes. Please ensure shoes are clear of mud, dirt, salt, rocks and debris before entering the Data Centers.

Overview of Computing Sector facilities and services

- We upgrade computing facilities as experiments' needs increase and as new technologies and budgets allow. Today's GCC tour shows a "snapshot" in time: January 2013.
- The Computing Sector (CS) operates nine computer rooms and three communications rooms at Fermilab that include over 30,000 square feet of space for computing equipment.
- CS provides email, web, database, business systems and disk servers, networking, and tape robotic storage. This also includes high density computational farms and high performance computing systems.
- CS manages more than 6500 computers (multi-CPU, multi-Core), greater than 16 PetaByte of disk storage (about 90% of which is for CMS) and 7 tape robots that in total consume over 3.5 megawatts (MW) of power. GCC alone consumes 2.25 MW of power.
- CS has 40 PB of data as of October 2012 (in FCC and GCC) robotic storage
- 2 WAN connections from FNAL (for redundancy):
    - One from GCC to LCC, through other points onsite, then connects off-site via connections near Giese and Kirk Roads to Argonne
    - One from FCC to Northwestern U's Starlight hub, a 1GigE and 10GigE switch/router facility for high-performance access to participating networks

Overview of Grid Computing Center facilities and services

- The Grid Computing Center (GCC) is sited at the WideBand service building, a former experimental hall with substantial power infrastructure.
- The building was built/adapted for GCC use in phases starting in 2004.
- Currently: 10,384 sq ft of raised floor.
    - We used raised floor to facilitate air flow (cooling, humidity, etc.).
    - All power and network cables are suspended above the equipment.
- The Grid Computing Center houses state of the art computing systems that support the Fermilab Scientific Experiments and program. There are four parts to these:
    - The Tape Robots which receive scientific data from the experiments data acquisition systems, and archive the processed data for ongoing analysis.
    - The Networking infrastructure which brings 10s of Gigabits/second of data to and from the building and between the systems in the building.
    - The Compute Farms (Worker nodes and server computers) which process the scientific event data for CMS, CDF and DZero RunII, and more.
    - The on-line disk storage arrays used for long-term storage as well as data staging for batch processing.

These are all served by a significant complex of grid-based electrical power, air-conditioning and cooling, battery and UPS backups.

Additional Facts, Trivia and Frequently Asked Questions

This portion of the document will probably change more often than the prior sections. As new questions are recognized as "Frequently Asked" or new interesting Facts are added.

Tape Robot Room

- Each robot with tape drives costs about $750K
- Currently, we have 167 tape drives + movers,in GCC and FCC with a total potential bandwidth of 27 GB/s
- All power to the robots, tape drives and movers is redundant. They are hooked up to a 40 KVA (kilovolt-Ampere) UPS (uninterruptible power supply) in the TRR and also the 1000 KVA UPS in CRA. This can maintain the state of the robots for a few hours, long enough to allow for a “soft” shutdown and/or obtain and hook up an external generator.

Network Rooms (A & B)

- Fact: There are approximately X miles of ethernet cables installed in the Grid Computing Center.
- The workbenches are for staff to bring a computer out of the unfriendly computer room environment into a quieter area for simple repairs or diagnostics.

Electrical Rooms

- Fact:

Computer Room A

- Fact:

Computer Room B

- Fact:

Computer Room C

- Fact: The HPC clusters are named Jpsi, Ds (D sub S) and Bc (B sub C). These are names of three different types of stable meson particles.

Tape Robot Room (TRR)

The TRR was built in first phase in 2004 along with Network Room A (NRA) and Computer Room A (CRA).

Clean Room

- These high capacity tapes and high precision tape drives are very sensitive to temperature and humidity changes and also to airborne contaminants.
- To help control the environment around the robots, visitors/tours view this room from the hallway, through the glass wall.

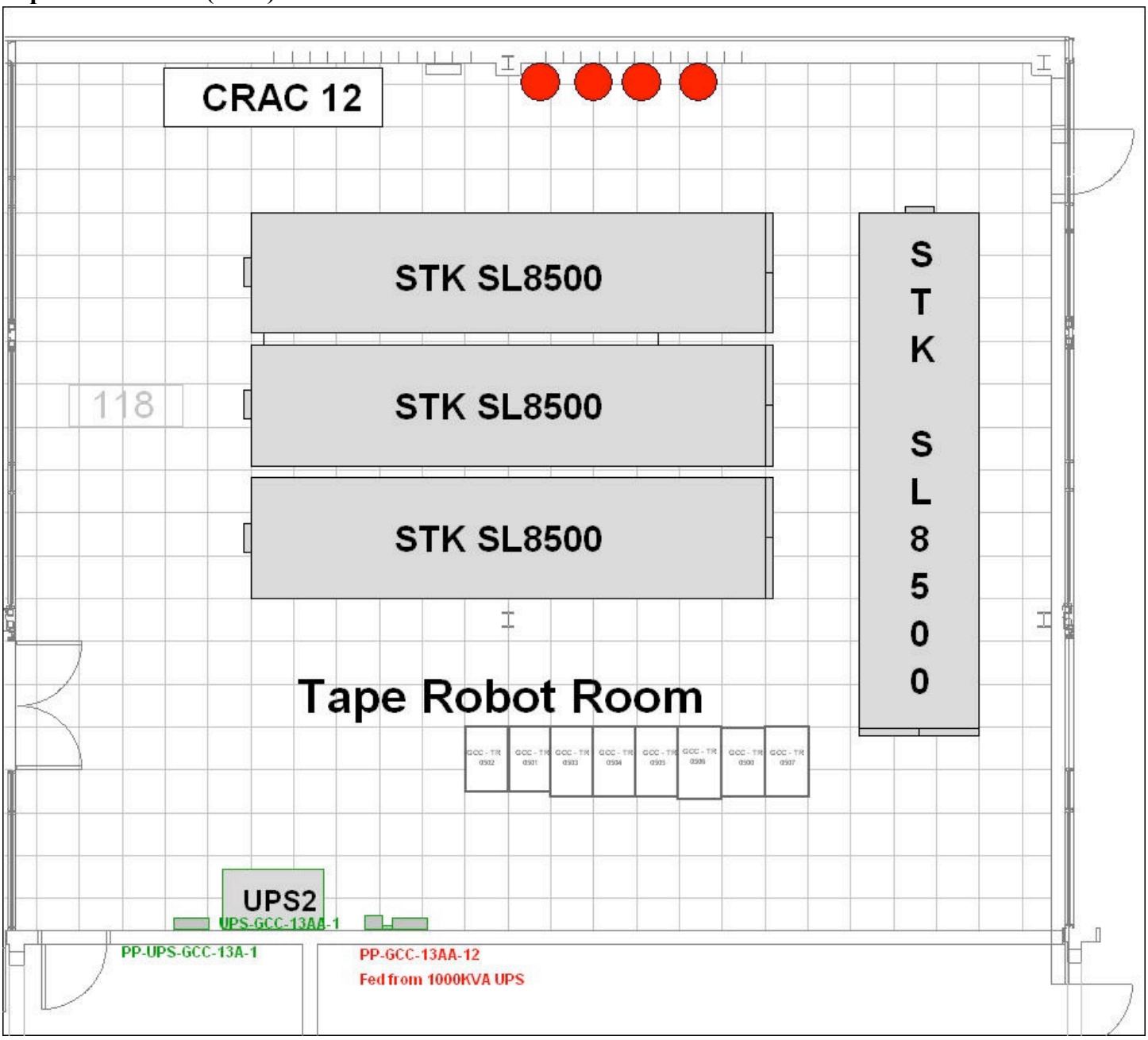
About the tape robots

- Four STK SL8500 tape robots at GCC hold 10,000 tapes each As of Oct 2012, over 38,000 slots were in use in these four robots.
- Tapes are a combination of 86% LTO4 (newer) and 14% T10000C media.
    - Each LTO4 tape holds 800 GB and transfers up to 120 MB/s
    - Each T10000C tape holds 5400 GB and transfers up to 240 MB/s
- Three of the four robots can pass tapes between each other via interconnecting "mailboxes" at the back of the robots. This triplet of robots serves the CMS experiment.
- "RAW" data from CDF and DZero experiments are stored in the rear stand-alone tape robot.
- The Computing Sector is in the process of migrating data on LTO4 tapes to T10000C tapes, which increases capacity, reduces space, and refreshes the technology.

"Follow the data"

- Purple cables (category 6, standard for Gigabit Ethernet) carry user requests for data from outside, through NRA, to "mover nodes" in the black racks.
- Mover nodes get assigned the user requests and ask the tape library to mount the appropriate tape.
- The robot finds the tape, and loads it into one of the tape drives associated with the mover.
- For reads, the mover pulls the data off of the tape drive (via the 2 or 4Gb/s orange fiber optic cables) and transmits it over the network to the user at transfer rate of up to 120 MB/s. For writes the data flow is the inverse.
- Between the mover and NRA, the data goes over the purple cables.

Tape Robot Room (TRR)



Floor plan updated January 31, 2013

Network Room A (NRA)

General Info

- Started operation in first phase, in late 2004.
- The cable management rails are suspended from the ceilings in all the rooms. In general, the copper-based ethernet cables follow the rails and fiber optic cables are routed through separate yellow cable management that is suspended beside the rails or beside electrical feeds.
- Note the red bags stacked above the cables where they move from one room to the next. These are fire retardant bags used as a firewall between rooms. These red bags were tightly tucked into place after the cables were first run between the rooms.

Connections

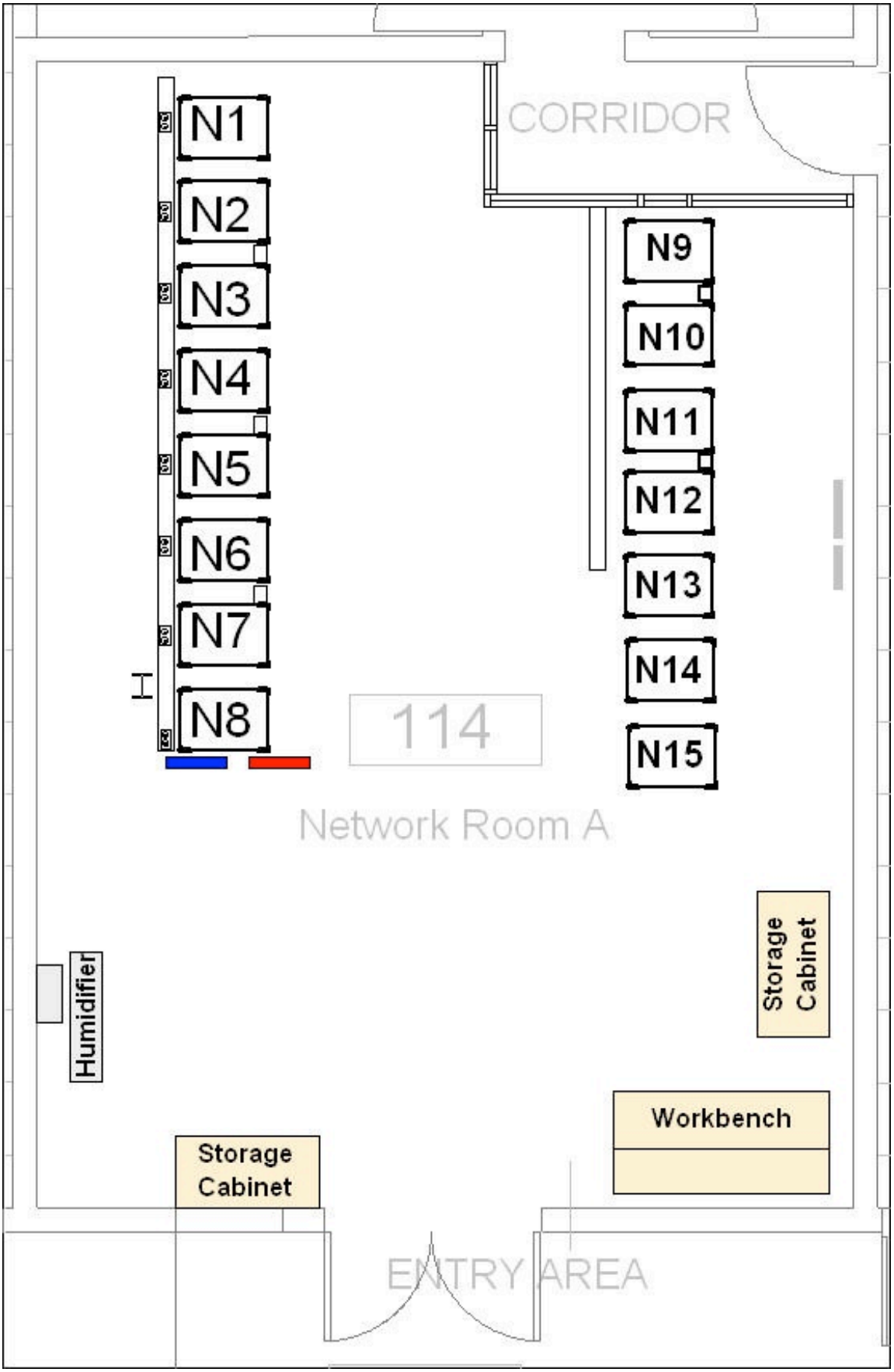
- Racks N1 – N8 (to the left as you enter room) are patch panels for the (currently about 3000) computers in Computer Room A (CRA), connecting them to the Internet.
- All connections to FCC happen here in NRA from rack N8 (yellow cables):
    - 144 pair capacity (currently about 80 used) of single mode[1] fiber (288 strands) between GCC and FCC
    - Most are 10GbE links, where "10GbE" is 10 times faster than Gigabit Ethernet
    - Top panels in N8 for CRA, next two down for CRB/C, coming from NRB
    - Also capability for multi-mode[2] fiber (aqua-colored cables); not currently in use
- In addition to connections to FCC, many connections here are between different computers in GCC and between computers and tape robots.
- Blue cables to right of panels (e.g., on N4) are the admin channels, used for reboots, etc.
- N1 can accommodate 1200 cables strung to/from CRA. 1000 used that are patched in N2 to connect to (N8 then to) FCC.
- Facing the back of the room, in upper left-hand corner, see black cable with the 144 pairs of fiber connecting GCC to FCC
- On right as you enter room, the black rack N9 holds network switch and patch panels to connect the purple cables from the mover nodes in TRR to the Internet.
- N10 routes general traffic to FCC and WH
- N13 provides WAN connection to Argonne (one of two paths for outside connections at FNAL)

---

[1] Single *transmission* mode; it carries higher bandwidth than multi-mode fiber, but requires a light source with a narrow spectral width; capable of higher transmission rate and up to 50 times more distance than multi-mode.

[2] Multi-mode: light waves are dispersed into numerous paths, or modes, as they travel through the cable's core; high bandwidth at high speeds (10 to 100MBS – Gigabit) over medium distances (275m to 2km))

Network Room A (NRA)



Floor plan updated March 14, 2013

Computer Room A (CRA)

General Info

- Started operation in first phase, in late 2004.
- About 3000 dual and quad-core computers (4 full rows, 2 half rows, 16 racks/row, 40 computers/rack)
- The computers run Linux, have 4 GB of memory and a CPU speed of 2 GHz
- Computers have been procured by various experimental groups (CDF, DZero, CMS) and by grid projects
    - In CRA: ~1600 for CMS, ~820 for DZero, ~470 for CDF, ~120 for grid
- All power and network cables that feed into the racks are suspended above the equipment.
- Each rack can handle 10 kW, therefore CRA can use up to about 800 kW in total (the load varies with processing activity)
- Data transfers over category 5 cable (fairly cheap) from/to NRA.

Environmental Controls

- This room has 9 CRAC (computer room air conditioning) units.
- Each CRAC unit is a device that monitors and maintains the temperature, air distribution and humidity levels in a network room or data center.
- All of the CRAC units can be remotely monitored by Facilities. There are also independent devices which are remotely monitored.
- We use raised floors to facilitate air flow.
    - Cold air is fed under the floor.
    - The cold air comes up through grates in the cold aisles.
    - The cold air is drawn into the equipment from the cold aisles and vented out the back into the warm aisles
    - Warm air is drawn back into CRAC units from above.
- Only a few of the CRAC units are used to add humidity into the room's environment.

How the experiments use these computers

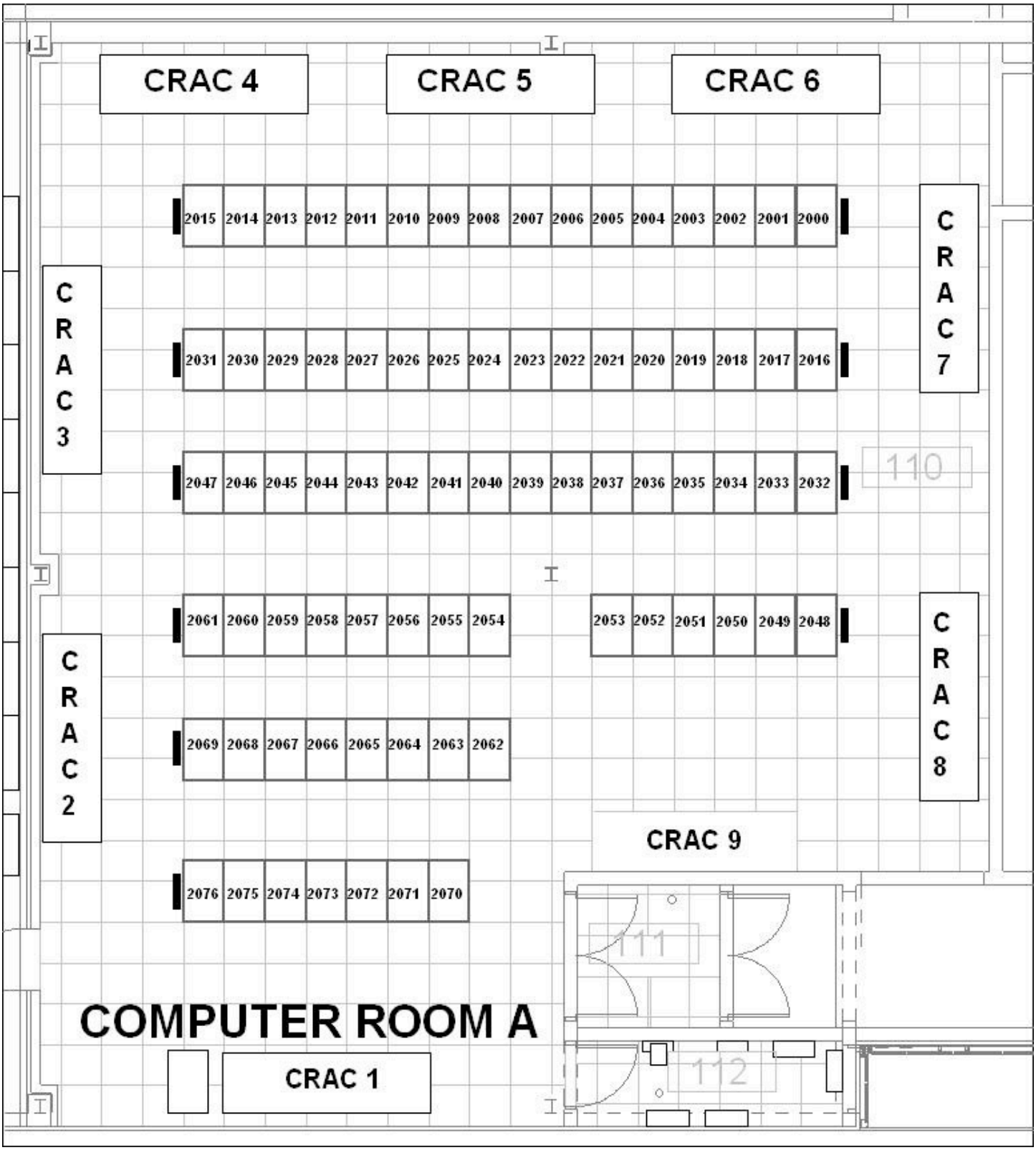
CDF and DZero use these nodes for both data reconstruction (turning raw data into physics objects) and physics analysis.

- Virtually all physics analysis is done on these nodes for both experiments.
- They are researching about 100 different physics topics
- The most exciting (and well-known) topic is Higgs search

For CMS: This is a Tier 1 center which runs data reprocessing, bulk analysis, simulation. To first order, all CMS servers and data are in FCC and all workers are in GCC:

- All FNAL-destined CMS data from the CERN Tier-0 comes in on the network at FCC, and is written to disks there.
- From there, the data on the disks are written to archival tape in the tape robot room in GCC.
- In parallel, these same data are delivered from the disks in FCC to the workers in GCC where they are analyzed.
- Data output of the worker nodes jobs is also sent to the disks in FCC, *and* onto archival tape in GCC.
- Therefore, all "active" data lives in FCC - it is the place that has the standby generator, and "technically", it is at the end of the CMS Data Acquisition System.

Computer Room A (CRA)



Floor plan updated January 31, 2013

Backup power (UPS) and Electrical rooms (2 rooms)

General Info

- These two rooms were the southern end of the original building in late 2004.

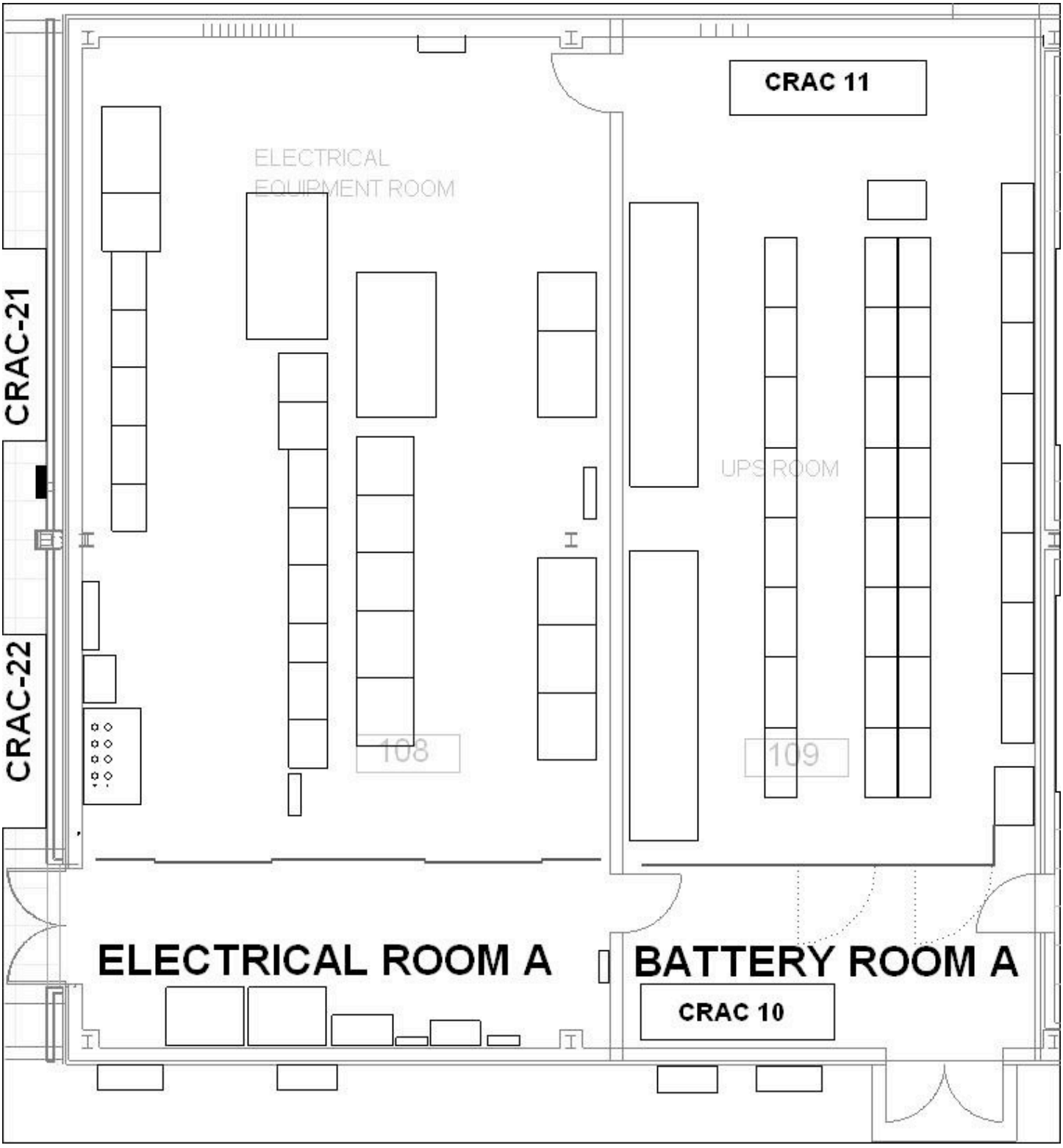
Backup power room

- GCC is powered by three underground 13.8 kV feeders from Fermilab's master substation. High voltage switches provide automatic switching from a failing feeder to a good one. Transformers, also external to the building, convert 13.8 kV to 480V.
- Tap boxes on exterior of building allow connection of portable generators to handle scheduled outages. Site outages typically occur once a year and 4-8 portable generators are rented.
- Three UPS units, each with 120 four-volt batteries, allow switching from the electrical power grid to rented generators and back. These newer batteries are "greener" than the previous generation which used 240 two volt batteries each.
- If the temperature in CRA, CRB or CRC reaches 125 degrees F (just under 52 degrees C), separate shunt breakers drop power to the room's computers.

Electrical room

- The electrical room can supply up to 2 MW for CRA and CRB.

Electrical and backup power (UPS) rooms



Floor plan updated January 31, 2013

Computer Room B (CRB)

General Info

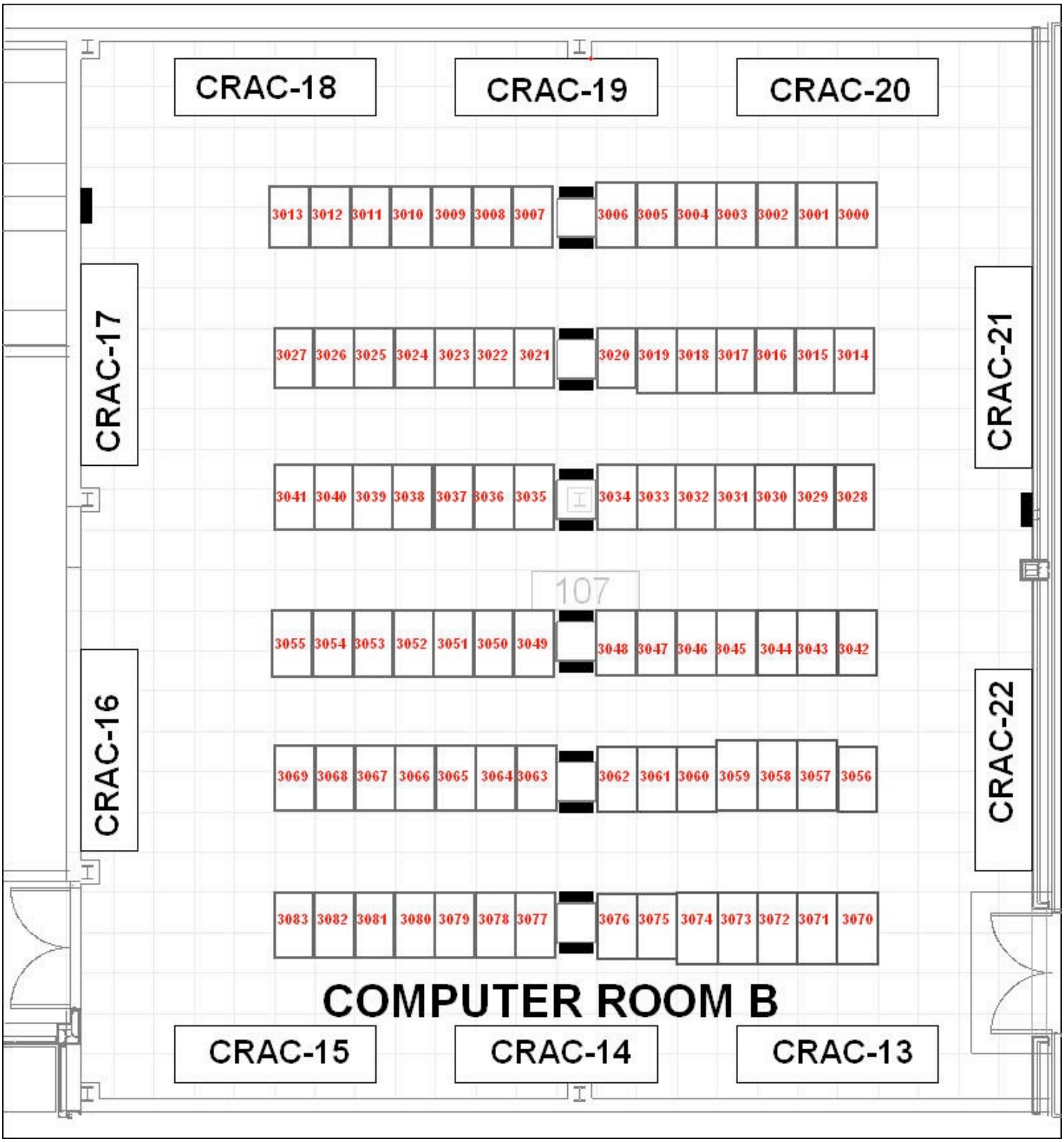
- CRB is in new addition south of the existing building; it was completed and started operation in early 2007.
- Like CRA, CRB was designed for 10kW racks; so 840 kW total capacity for CRB
- Similar computing capacity to CRA, but using fewer, faster computers:
    - Number of computers determined by cooling limit
    - Dual and quad core; minimal disk space
    - Most have 16GB of RAM or more; the latest have 24GB of RAM.
    - Clock speeds vary, many are more than 2.0 GHz
    - Over 2500 computers in CRB (6 rows, 14 racks/row, 30 computers/rack)
- CRB computers used for data analysis for CMS, DZero and CDF, also as grid resources, and some general purpose.
- The FermiCloud (described below) racks will be installed in GCC B shortly, there is one test FermiCloud rack there now.
- Data goes via Cat 5 or 6 cables to network room B; then switches to FCC via fiber in NRA.
- 10 CRAC (computer room air conditioning) units, remotely monitored and alarmed.

About FermiCloud

The FermiCloud project is working to deploy a flexible Infrastructure-as-a-service facility. It aims to deploy just-in-time build and test images for all supported Fermilab operating systems.

The scope of this project is to build a scientific testing and development private cloud at Fermilab. We will not deploy services on commercial clouds, although we will attempt to find a solution which is compatible with existing commercial clouds.

Computer Room B (CRB)



Floor plan updated January 31, 2013

Network Room B (NRB)

General Info

- Started operation in phase two (with CRB) in March 2007
- NRB has a rooftop air conditioner and in-room humidifier (as contrasted with the in-room AC units which have built-in humidifiers)
- NRB is remotely monitored and alarmed for temperature spikes

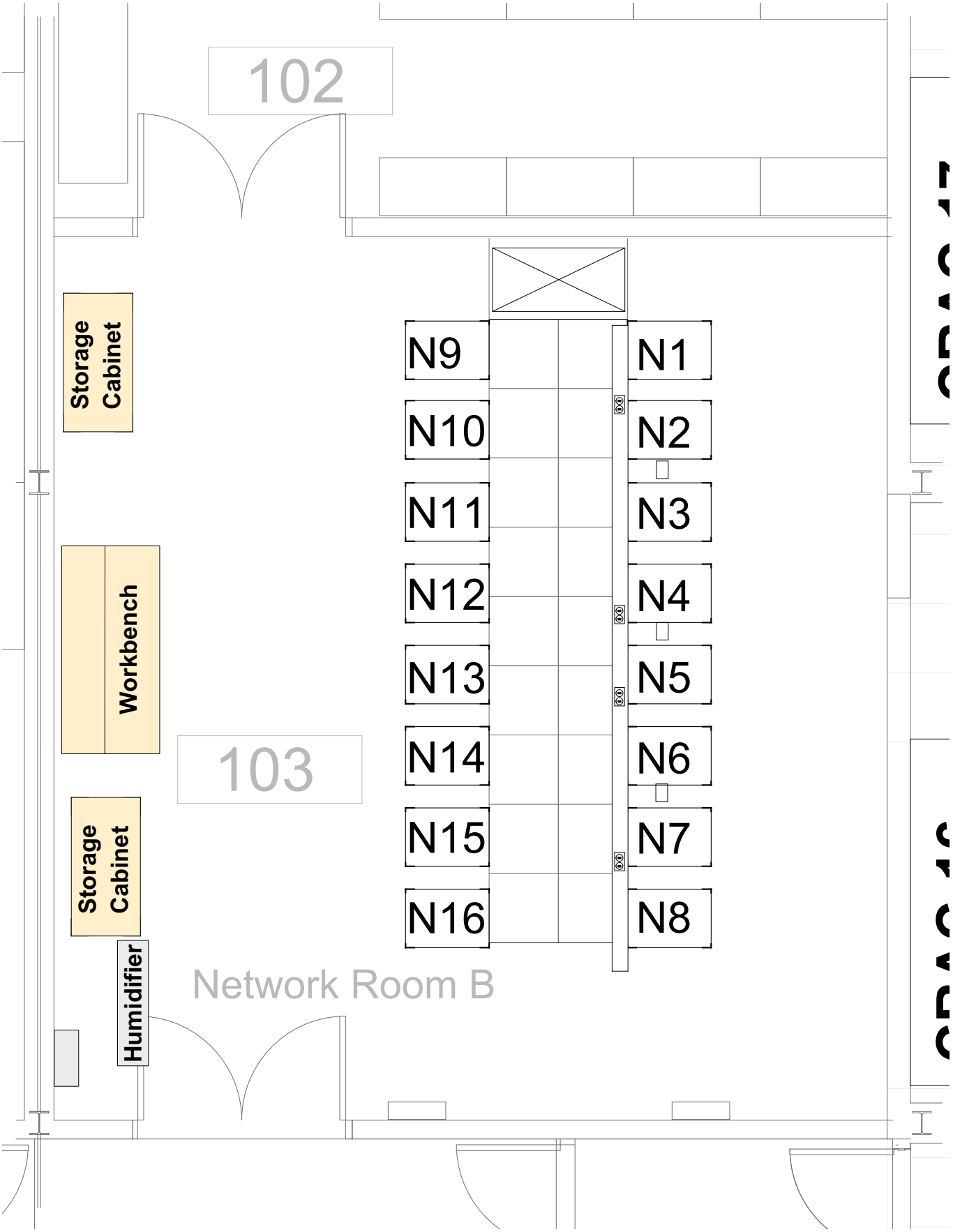
Connections

- Black racks provide patch panels for CRB computers, and 3 network switches for the three subsets of CRB computers: CMS, DZero and CDF
- Black cable with 144 pairs of fiber connecting NRB to NRA (no external connections from NRB)
- CMS network allocation: 200 GB/s (but they will need more)
- CDF and DZero each have allocation: 10 GB/s

Evolving Technology and Configurations

- When you look at the wall where network wiring passes into CRB, you see the familiar cables stacked on rails and packed with fire-retardant bags. There is one wire for each computer or device throughout CRB.
- When you look at the wall where network wiring passes into CRC, you see just a few fiber-optic cables running along the familiar rails.
- Each of those fiber connections goes to a network switch installed in racks inside CRC. The cables from those switches travel shorter distances to the computers and devices around the room.
- This allows for better network throughput due to localized and efficient switching.
- The other big benefit is we use much less copper based ethernet cables. And this simpler configuration was much easier to install.

Network Room B (NRB)



Floor plan updated March 14, 2013

Computer Room C (CRC)

General Info

- CRC added in summer of 2008
- Designed for 14kW racks (compared to 10kW racks for CRA and CRB) to accommodate faster computers.
- These racks are deeper than the older racks in the other rooms.
- 428 chassis, 856 nodes (two nodes per 1U chassis) (2 rows, 11 racks/row, 23 or 22 chassis/rack)
- Three GPU racks (details are included below)
- 208 V service, compared to 120V for the rest of GCC. (208V is 1-3% more efficient than 120V.)
- Systems in CRC are used for High Performance Computing, such as Lattice Gauge calculations. The HPC Department also manages clusters at LCC (Lattice Computing Center) for accelerator modeling and astrophysics calculations.
- Uses specialized Infiniband connections that connect nodes at the backplane. The newest QCD (Quad Data Rate) interconnects are capable of 40 GB/s bi-directional data transfers: the equivalent of 10 DVDs per second.
- 10 CRAC (computer room air conditioning) units, remotely monitored and alarmed.

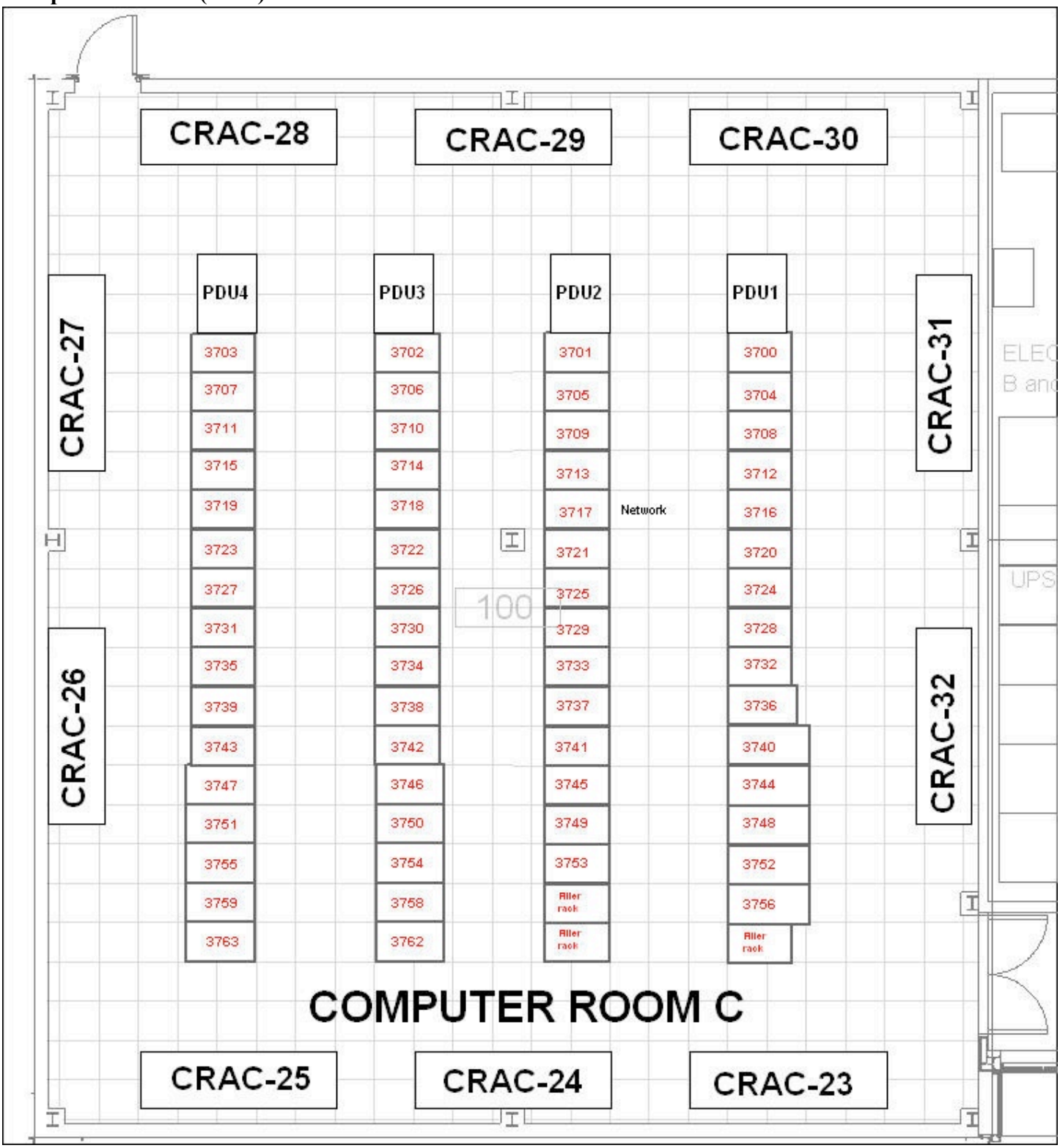
Storage Info

- Majority of storage in CRC uses Lustre (disk-based) filesystems. In addition, dCache (tape data staged to disk) and Enstore (tape) based storage is used.
- Lattice QCD Lustre storage is located here and data is transferred over 40Gb/s Infiniband over a private network. The Lustre storage uses hardware RAID arrays. These protect data integrity against failures of component disk drives within the arrays.
- Enstore and tape backed storage is accessed via Gigabit Ethernet (GbE) using category 5/6 cables to NRB, and sent on to FCC via fiber.

About GPUs

- (Adapted from Wikipedia) A graphics processing unit or GPU is a specialized processor that offloads 3D graphics rendering from the microprocessor. Modern GPUs are very efficient at manipulating computer graphics, and their highly parallel structure makes them more effective than general-purpose CPUs for a range of complex algorithms.
- The HPC department originally operated six nVIDIA Tesla S1070 GPUs: four for the Lattice QCD project and two for the Computing Sector. Each S1070 Tesla has four 1TFlop GPU processors.
- The six S1070 Tesla GPUs deliver a peak performance of 24TFlops vs. the jpsi cluster which delivers a peak performance of 56TFlops. In other words, six of these S1070 Teslas are equivalent to four jpsi racks.
- In January 2012, HPC added 5 racks of new Hewlett Packard GPU-based servers. There are 76 nodes each with dual Xeon 2.53 GHz quad-core CPUs, 2 nVIDIA Telsa M2050 3Gb GPUs, QDR Infiniband, two 1 GbE and 48 Gb of memory.
- HPC is currently working to migrate code run on CPU based clusters to run on Gen Purpose GPUs and realize higher performance floating-point calculations for simulations.

Computer Room C (CRC)



Floor plan updated January 31, 2013